

Mapping the Whole Human Genome by Fingerprinting Yeast Artificial Chromosomes

Christine Bellanné-Chantelot,* Bruno Lacroix,*
Pierre Ougen,* Alain Billault,* Sandrine Beauvils,†
Stéphane Bertrand,† Isabelle Georges,†
Fabrice Gilbert,† Isabelle Gros,†
Georges Lucotte,† Laurent Susini,†
Jean-Jacques Codani,‡ Philippe Gesnoux,†
Stuart Pook,† Guy Vaysseix,* Jennifer Lu-Kuo,§
Thomas Ried,§ David Ward,§ Ilya Chumakov,*
Denis Le Paslier,* Emmanuel Barillot,†
and Daniel Cohen*†

*Centre d'Etude du Polymorphisme Humain
75010 Paris

France

†Généthon

91000 Evry

France

‡Institut National de Recherche en Informatique
et Automatique

78153 Le Chesnay

France

§Departments of Genetics, Molecular Biophysics
and Biochemistry

Yale University School of Medicine

New Haven, Connecticut 06510

Summary

Physical mapping of the human genome has until now been envisioned through single chromosome strategies. We demonstrate that by using large insert yeast artificial chromosomes (YACs) a whole genome approach becomes feasible. YACs (22,000) of 810 kb mean size (5 genome equivalents) have been fingerprinted to obtain individual patterns of restriction fragments detected by a LINE-1 (L1) probe. More than 1000 contigs were assembled. Ten randomly chosen contigs were validated by metaphase chromosome fluorescence in situ hybridization, as well as by analyzing the inter-Alu PCR patterns of their constituent YACs. We estimate that 15% to 20% of the human genome, mainly the L1-rich regions, is already covered with contigs larger than 3 Mb.

Introduction

Mapping the whole human genome has become one of the major challenges of modern genetics. Mapping consists generally of the classification of genomic DNA fragments according to their order along the chromosomes. This order can be genetically determined when the fragments correspond to polymorphic sites between which meiotic recombination frequencies can be estimated. This is linkage mapping, indispensable to localizing any polymorphic or pathological trait gene or factor (Botstein et al., 1980). Genomic fragments can also be physically ordered using several methods, e.g., fluorescence in situ hybridization

(FISH) (Montanaro et al., 1991; Korenberg et al., 1992), somatic cell hybrids (Cox et al., 1990), or fingerprinting (Coulson et al., 1986; Olson et al., 1986; Kohara et al., 1987; Evans and Lewis, 1989; Craig et al., 1990; Stallings et al., 1990). The latter approach has been used to establish collections of overlapping genomic fragments called "contigs" covering all or part of the genomes of *Caenorhabditis elegans*, yeast, *Escherichia coli*, herpes simplex virus, and humans. A contig map covering the human genome will be of the utmost importance to accelerate the identification of hereditary disease genes. Indeed this map will provide immediate access to the genomic portion including any disease gene as soon as its locus is flanked with polymorphic markers by genetic linkage analysis. Only a few percent of the human genome has been covered by contigs, ranging from 200 to 2000 kb.

Until now, attempts of physical mapping have only been envisioned on single chromosomes. This was understandable given the tools available until the recent past. New technological improvements allow at present to forecast a significant acceleration of the process. This impulsion can mainly be attributed to two changes of scale: first, in the size of the cloned fragments and, second, in the fingerprint production throughout. Yeast artificial chromosome (YAC) technology permitted an increase in genomic cloning size by 5- to 10-fold over cosmids (Burke et al., 1987; Anand et al., 1989; Albertsen et al., 1990; Larin et al., 1991). We have recently been able to increase this further to a factor of 20 to 25 (P. Ougen, personal communication). From this improvement, it follows that the same effort would be required to cover the whole human genome with YACs or a single human chromosome with cosmids.

Such a whole genome mapping approach would require a priori fingerprinting of 30 to 50,000 YACs of 1000 kb mean size. This could not be achieved without a robust fingerprinting methodology that is based here on the ability to obtain the patterns of restriction fragments carrying repeated sequences (Cangiano et al., 1990; Stallings et al., 1990; Bellanné-Chantelot et al., 1991); for this purpose, the process has been largely automated.

We report contigs assembled from 22,000 random YACs fingerprinted using a LINE-1 (L1) probe (Shafit-Zagardo et al., 1982). Moreover, we present data showing that this approach will quickly permit covering more than 90% of the whole human genome.

Results and Discussion

YAC Fingerprinting and Contig Assembly Procedures

Yeast colonies were grown in 96-well plates and subjected to DNA extraction by a simple lysis-dialysis procedure. DNA samples were then digested with EcoRI, PstI, and PvuII, chosen for their reliability. Blotting is automatically performed by a device that handles loading, migration, and transfer. Membranes are then hybridized with a chemiluminescent L1 probe. Autoradiograms are digitized

and interpreted by automatic image analysis software (Figure 1). At this stage the fingerprint consists of three series of interpolated fragment sizes, one for each enzyme.

The contig assembly is achieved in two steps. The first performs pairwise comparisons of all the fingerprinted YACs. Each pair is assigned an overlap likelihood value using Bayesian statistics according to a model previously described (Balding and Torney, 1991; Lacroix and Codani, 1992). For each pair of YACs, the most probable length of physical overlap is estimated from the fingerprints (see Experimental Procedures). The second step consists of forming contigs by selecting all the pairs determined above a given likelihood threshold value. The contigs expand when assembling those pairs sharing 1 YAC. Once a large contig is formed, ordering the YACs within it remains delicate and algorithms are being developed to determine internal order. Finally, any contig can be positioned on metaphase chromosomes using FISH with a probe obtained by pooling the inter-Alu polymerase chain reaction (PCR) products of its YACs.

Evaluation of Experimental and Biological Parameters That Can Affect the Contig Assembly Process and Internal Ordering

The fingerprinting process was tested for partial digestions, error frequencies in band detection (false positives and false negatives), and standard deviation of fragment size measurement.

Completion of digestion was assessed by hybridization with two yeast single-copy probes, pAF001 and pAF020, for each of the three enzymes. A total of 402 independent DNA preparations were tested and gave 1, 2, and 0 partial digestions for EcoRI, PstI, and PvuII, respectively. This rate, systematically monitored on 1% of the samples, is low enough to have little effect on contig assembly. The other experimental parameters were estimated by analyzing data from 24 independent preparations and migrations of 4 YACs of various sizes (300, 400, 660, and 950 kb) digested with the three enzymes. The analysis of the resulting data base gave the distribution of standard deviations of size measurement. This distribution was found to be Gaussian and varies with the fragment length. The relative standard deviation increases from 0.3% for 1 kb fragments to 1.7% for 20 kb fragments. This function was introduced in the algorithm, calculating the likelihood of overlap score values, designated LOS values. We also used this data base to obtain information on false positive (that includes partially digested fragments) and false negative band frequencies. In summary, a single enzyme fingerprint out of 24 has an artifactual band (false positive). No fingerprint has more than one artifactual band. Eighty-three percent of the fragments are reproducibly detected, all of which have an optical density greater than 0.2. The reproducibility of detection drops when optical densities fall below this value.

Genetic polymorphism can affect the pairwise comparison since the YAC library was constructed from a diploid genome (Albertsen et al., 1990). The incidence of restriction fragment length polymorphism on fingerprint patterns can be estimated, assuming that 1 base out of 300 is af-

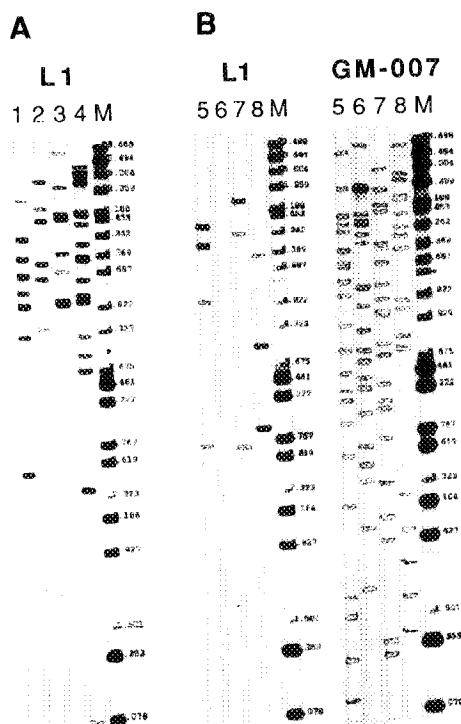


Figure 1. Digitized Images of Autoradiograms Obtained with L1 and GM-007 Probes

YAC clones were digested with PvuII and hybridized with chemiluminescent L1 and GM-007 probes. YAC clones in (A) are L1 rich, whereas YACs in (B) are L1 poor or L1 negative. The probe GM-007, derived from the Alu consensus sequence, was tested on the L1-poor clones ([B], lanes 5, 7, and 8) and on a L1-negative clone ([B], lane 6). The size marker is a mixture of λ DNA cut with XhoI, BssHII, or BstEII and ϕ X174 cut with HaeIII, MluI, BstNI, or NruI.

ected by a point mutation and that about 2000 variable number tandem repeats exist in the human genome. Point mutations should then eliminate 1 restriction site in 50 and create 1 every 200 kb. The two haploid genomes will, therefore, only differ for approximately 8% of their restriction fragments (4% for fragments losing one boundary site and 4% for fragments gaining one internal site, with the mean L1-positive fragment size being 8 kb according to our data). Thus, because of restriction fragment length polymorphisms, the fingerprints of two clones covering the same portion of the paternal and maternal genome, respectively, could rarely differ by more than one fragment. It is still difficult to evaluate the effect of variable number tandem repeats. If they contain no restriction site, their frequency being 1 every 1600 kb, they should only affect 0.5% of the L1-positive fragments, a negligible proportion. But this rate could vary drastically according to the regions; variable number tandem repeats are not uniformly distributed on the human genome, but are concentrated in subtelomeric regions (Jeffreys, 1987).

Approximately 40% of the YACs from this library are chimeric, i.e., they contain inserts with more than one genomic origin (P. Ougen, personal communication). This is either due to coligation during YAC construction or recom-

bination events between cotransforming DNA fragments (Green et al., 1991). They represent an essential problem for physical mapping. These clones are less efficiently integrated into contigs, but can also create false contigs by merging two subcontigs with different genome origins. Moreover, when included in a genuine contig they complicate the internal clone ordering by introducing nonmatching fragments that create ambiguities. Actually, such nonmatching fragments may also be found in YACs located at contig extremities, or can correspond to partial digests, to novel fragments created at the insert vector arm junctions, or to polymorphic fragments.

The rate of colonies with more than 1 YAC was estimated to be 2.3% (P. Ougen, personal communication). Most of them contain two or more rearrangements of a single initial YAC (Bates et al., 1992), and such rearranged forms will have no effect on the mapping process. It must also be pointed out that homologous regions dispersed in the genome will generate chimeric contigs.

Finally the genomic distribution of the repeated sequence will affect the coverage and intuitively a uniform distribution would be optimal. Actually, it has been reported that the L1 sequence is more frequent in genomic regions corresponding to metaphase G bands (Korenberg and Rykowski, 1988). Therefore, the contigs we report will preferentially cover these regions.

Contig Assembly of 22,000 YACs of 810 kb Average Size (5 Genome Equivalents)

A total of 22,000 random YACs of 810 kb average size (Figure 2) were fingerprinted using three enzymes, EcoRI, PstI, and PvuII, and an L1 probe. All pairwise comparisons were performed, and the corresponding LOS values were computed. For practical reasons, the LOS values are stored as $7(\log R)$, where R is the ratio of the probability of observing the two patterns assuming overlap versus nonoverlap. Overlapping pairs are more likely to yield a large LOS, while nonoverlapping ones yield a small LOS. The LOS is not the a posteriori probability of overlap according to the data, which is difficult to compute for two reasons: first, the precise size of each clone is unknown, and, second, the equation, which takes into account the interrelationship of the fingerprints obtained with the three enzymes, is intractable. In practice, a synthetic LOS is calculated, considering that the three fingerprints are independent, by taking the mean LOS value for the three enzymes. Consequently, all the pairs characterized by a LOS value above a certain threshold are selected and then assembled into larger contigs.

The determination of the threshold value is obviously a critical step. Some simulations indicated that a threshold around 60 would be appropriate. To confirm this value, we analyzed the LOS value of 18,642 nonoverlapping pairs derived from an ongoing contig map on chromosome 21 (I. Chumakov, personal communication). Such nonoverlapping pairs were obtained by using fingerprinted YACs mapped to different metaphase bands of chromosome 21. This analysis suggests that a study of 22,000 random YACs would retain only 10 false pairs when choosing a LOS threshold of 60 (Table 1).

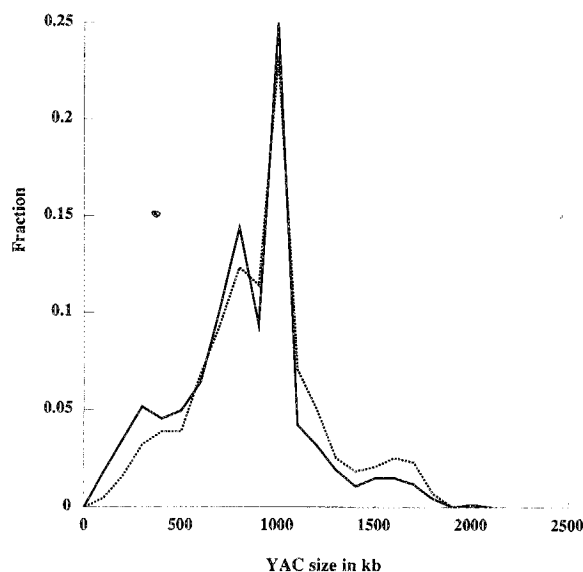


Figure 2. YAC Size Distribution

This was established from 953 clone sizings for the 21,696 fingerprinted YACs (solid line) and from 437 clone sizings for the 6,621 YACs assembled in contigs at the threshold 60 (dashed line).

Contig assembly was performed with LOS thresholds from 58 to 64 (Table 1). The number of contigs varies slightly between 1175 and 1241, while, as expected, the number of clones in contigs increases as the threshold decreases. The mean number of clones per contig varies from 4.3 to 5.7. Nearly two-thirds of the clones remain isolated. Actually, we compared each clone to itself to uncover those that could never exceed the threshold, i.e., are noninformative. The fingerprint of these clones, roughly one-third, is in fact too sparse. This implies that other probes are required to incorporate such clones into contigs (see below). Size distribution of the contigs measured as the number of clones per contig is also shown. By lowering the threshold, the predicted number of YACs incorrectly incorporated into contigs increases as previously described from 1 to 52 (Table 1).

Checking the Validity of Contig Assembly

We selected 10 contigs with a LOS value ranging from 58 to 94 (Table 2). Their sizes varied between 4 and 21 YACs. They were analyzed simultaneously by metaphase chromosome FISH and by comparing the inter-Alu PCR patterns of their respective YACs. The metaphase chromosome FISH was performed using pooled inter-Alu PCR products from the constituent YACs as probes. For six contigs of 4, 8, 10, 14, 18, and 21 YACs, respectively, a single chromosomal location was found, thus validating these contigs (Figure 3A). The inter-Alu PCR fragment size patterns obtained from the corresponding YACs corroborates the validity of these contigs since in most cases every YAC in these contigs shared at least one band with at least one other member of the same contig. The computer-reconstituted L1 fingerprints (1A, 1B, and 1C) and the inter-Alu PCR patterns (1D) of one of these contigs (number 2)

Table 1. Results of the Assembly Process

| Numbers of Clones and Contigs | Threshold | | | |
|-------------------------------------|-----------|--------|--------|--------|
| | 64 | 62 | 60 | 58 |
| Number of clones in contigs | 5,091 | 5,638 | 6,222 | 6,897 |
| Number of isolated clones | 16,605 | 16,058 | 15,474 | 14,799 |
| Number of contigs | 1,175 | 1,229 | 1,241 | 1,205 |
| Mean number of clones per contig | 4.3 | 4.6 | 5.0 | 5.7 |
| Number of noninformative clones | 7,937 | 7,482 | 7,141 | 6,798 |
| Predicted number of false positives | 1 | 3 | 13 | 52 |
| Number of contigs with | | | | |
| 2 clones | 530 | 558 | 523 | 486 |
| 3 clones | 199 | 191 | 197 | 187 |
| 4 clones | 105 | 117 | 128 | 118 |
| 5 clones | 86 | 77 | 75 | 79 |
| 6 clones | 64 | 62 | 64 | 55 |
| 7 clones | 44 | 54 | 57 | 46 |
| 8 clones | 28 | 31 | 30 | 38 |
| 9 clones | 30 | 30 | 30 | 36 |
| 10 clones | 19 | 18 | 26 | 27 |
| 11-15 clones | 37 | 49 | 53 | 58 |
| 16-20 clones | 13 | 14 | 26 | 34 |
| 21-30 clones | 18 | 22 | 20 | 19 |
| 31-50 clones | 2 | 6 | 9 | 12 |
| 51-100 clones | 0 | 0 | 3 | 10 |

are shown in Figure 4. For each of the four remaining contigs, two locations were found. This could be due to wrong contig assembly, to chimeric YACs, or to assembly of clones from unlinked homologous regions. In three cases (contigs number 4, number 9, and number 10), the inter-Alu PCR pattern indicates evidence for good contig assembly. In the last case (contig number 7), two YACs remained unlinked to the others, but this was also observed with members of some contigs assigned to a single location by FISH. Moreover, for contig number 7 and number 10, comprising 10 and 4 YACs, respectively, YACs were checked individually by FISH. Contig number 7 was assigned to chromosomes 1q24-25 and 10p11.2-12. Three YACs were mapped on 1q24-25 (Figure 3B), four on 10p11.2-12 (Figure 3C), and one on both chromosomes (Figure 3D). At this stage, the data suggest chimerism, but an additional chromosome assignment for two of the contig clones indicates homology since one mapped on 1q24-25, Xp11.3, and 7q36 (Figure 3E) and the other mapped on 10p11.2 and Xp11.3 (Figure 3F). This indicates

that regions 10p11.2-12, 1q24-25, and Xp11.3 present strong homology. Contig number 10 was assigned to 8q23 and 8q21. Three YACs hybridize only to 8q23, and the last one hybridizes to 8q23 and 8q21, either by chimerism or by homology. Finally, for contigs number 4 and number 9, we cannot yet discriminate between chimeric YACs and unlinked homologous regions. It is intriguing that four contigs (numbers 5, 6, 9, and 10) were mapped on chromosome 8 and that in two cases (numbers 9 and 10), where two locations were found, the second was also on chromosome 8.

At a LOS value of 58, contig number 3 increases from 14 to 34 members. For only one of these clones (681H7), the inter-Alu PCR pattern (2D in Figure 4) does not match, as shown together with the L1 fingerprint (2A, 2B, and 2C in Figure 4). Figure 4 also indicates the size of the YACs merged in this assembly process. All the YACs of the ten contigs were sized, and the mean size of YACs in contigs is significantly larger than the mean YAC size of the starting library shown in Figure 2 (900 kb versus 810 kb). Indeed, larger YACs are expected to express more matching information.

Taken together, these data suggest that most of these contigs were properly assembled. The fact that chimerism did not cause a major problem in contig assembly can, probably, be attributed to the fact that at the threshold of 60, most of the pairs overlap over a large portion of their length. Indeed, from all our data we have estimated that at this threshold, 70% of the assembled pairs share at least 60% of their length. This tends to prevent chimeras from creating false contigs.

It is possible to impose a given minimal physical overlap to each pair selected at a given threshold. Assembling contigs with such pairs will minimize the number of chimeric contigs generated by chimeric or multiple clones.

Table 2. Checking Ten contigs by FISH and Inter-Alu PCR Patterns

| Contig | Number of YACs | LOS | Location | Number of YACs Not Sharing Inter-Alu PCR Fragments |
|--------|----------------|-----|------------------|--|
| 1 | 4 | 94 | 4q31 | 0 |
| 2 | 8 | 64 | 5q11.1 | 0 |
| 3 | 14 | 62 | 5q13 | 1 |
| 4 | 5 | 78 | 2q35/3p26 | 0 |
| 5 | 18 | 59 | 8q11.2 | 1 |
| 6 | 21 | 61 | 8p11.2, 8p12 | 2 |
| 7 | 10 | 58 | 1q31/10p11.2-p12 | 2 |
| 8 | 10 | 64 | 2q33 | 1 |
| 9 | 11 | 63 | 8p11.2/8q11.2 | 0 |
| 10 | 4 | 66 | 8q23 (8q21) | 0 |

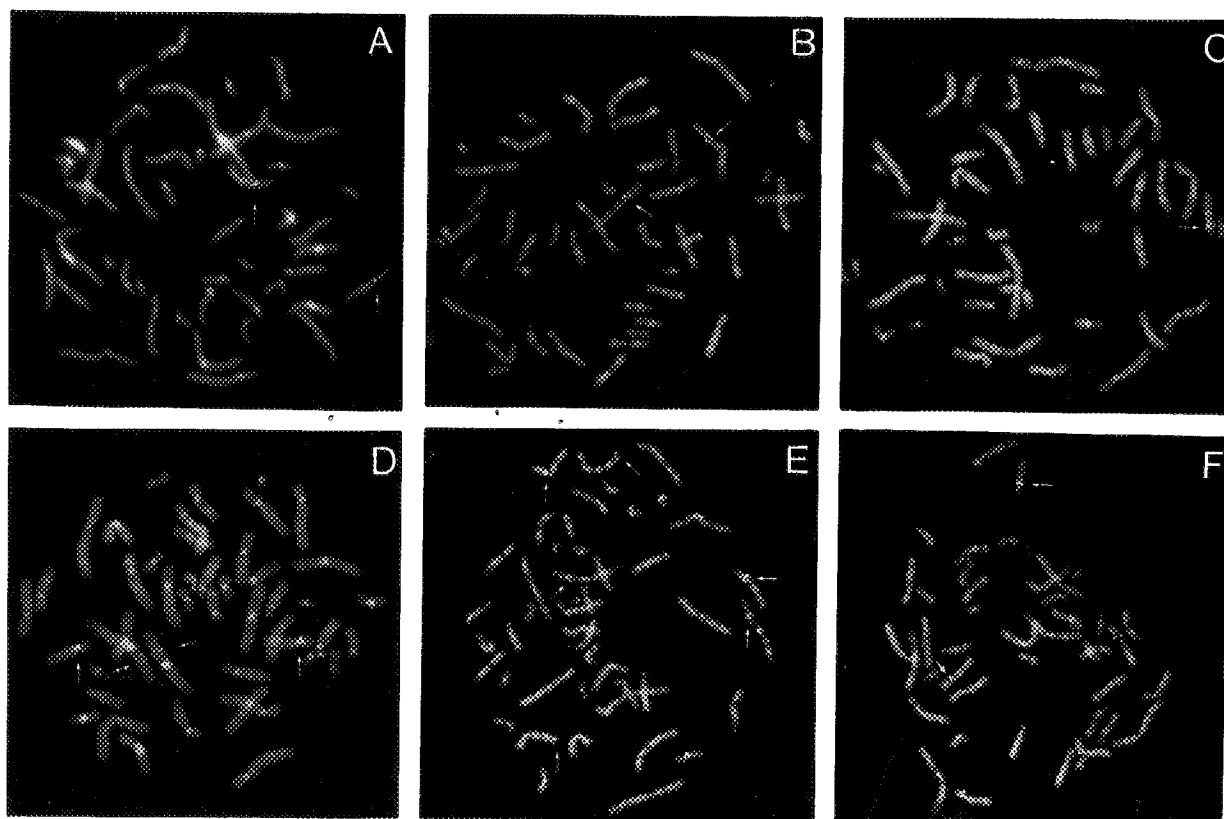


Figure 3. FISH Mapping of YAC Contigs

Metaphase chromosome spreads showing hybridization signals (arrows) generated by biotinylated Alu PCR products amplified from a pool of 18 YACs constituting contig 5 (A), map position 8q11.2) and individual YAC clones from contig 7 (B-F). (B), clone 703A11, 1q24-25; (C), clone 759C11, 10q11.2-12; (D), clone 754D10, 10p12.2-12, and 1q24-25; (E), clone 798B7, 1q24-25, Xp11.3, and 7q36; (F), clone 798C7, 10p11.2, and Xp11.3. Signals were detected with fluorescein-labeled avidin and chromosomes banded by DAPI staining.

This solution is being implemented at the present time. Another way to eliminate most illegitimate junctions created by these clones is to detect them beforehand, by assigning each YAC to its chromosome, i.e., screening the YAC library with probes representative of single chromosomes. Such probes have been successfully generated for chromosome 21 from a human-rodent somatic cell hybrid containing only that chromosome, by labeling the corresponding inter-Alu PCR products (Chumakov et al., 1992). Representative chromosome-specific subsets have also been obtained for larger chromosomes (I. Chumakov, personal communication).

The Proportion of the Human Genome Covered by This Mapping Strategy

This proportion largely depends on the genomic distribution of L1 repeats that conditions the clone information content. In fact, we observed that the mean number of fragments detected per YAC and per enzyme with this probe, which is 5 for the whole library, jumps to 8 when considering the YACs assembled into contigs. This increase cannot be accounted for by the previously mentioned bias toward longer clones (from 810 to 900 kb) in contigs, and implies that these contigs preferentially cover L1-rich regions, i.e., regions that have higher information

content. An a posteriori analysis of the L1 fingerprints of 953 randomly chosen and sized clones was performed assuming the existence of two types of regions, as suggested previously (Korenberg and Rykowski, 1988): L1-rich regions and L1-poor regions. An unexpected fraction of negative clones suggests the presence of a third type of region that is L1 negative. Supposing that in the two L1-positive regions the repeated sequence was uniformly distributed, we evaluated the repeat frequency and the abundance of each type of region: 10% L1 negative, 27% L1 rich (12 repeats per Mb), and 63% L1 poor (3.7 repeats per Mb). This model probably represents a rough approximation of the real genomic distribution. Nevertheless, this estimation was necessary to evaluate the effective coverage of the genome.

Equations have already been derived to predict the contig sizes (in Mb) as a function of the number of clones fingerprinted. These equations considered the clone size and minimal detectable overlap distributions (Lander and Waterman, 1988). New equations have been developed in order to integrate crucial parameters, such as the ratio of noninformative clones at a given LOS threshold, the chimera rate, and the hypothesized distribution of the repeated sequence used for fingerprinting. Parameters of these equations have been evaluated from simulated data



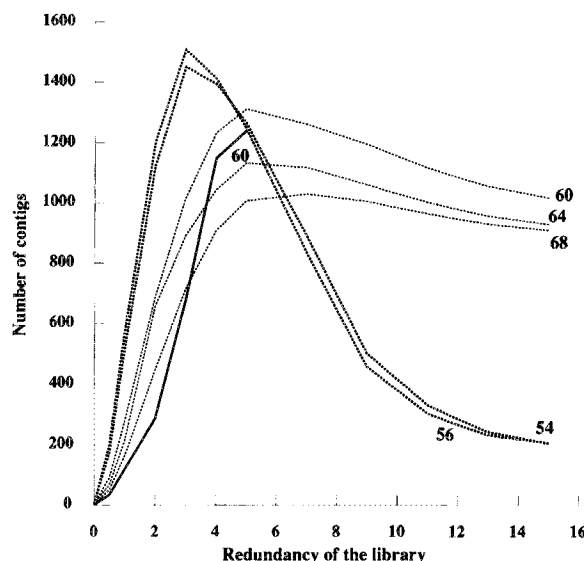


Figure 5. Observed and Predicted Results

Predicted numbers of contigs (dashed lines) are displayed as a function of the redundancy for various threshold values indicated, using either an L1 probe (thin dashed lines) or an L1 probe along with GM-007 (bold dashed lines). Observed results are displayed as a solid line.

on 10,000 clones covering 500 Mb, which were fingerprinted with three enzymes and probes that detect repeated sequences with various uniform genomic distributions. From this prediction we have plotted (Figure 5) the expected number of contigs against the redundancy of the library being fingerprinted. This was evaluated for LOS threshold values of 60, 64, and 68 with the L1 probe (assuming the trimodal distribution already described). The observed mapping progress for a threshold of 60 is superimposed and broadly fits expectations. We concluded that the results obtained with simulated data at a threshold of 64 were the closest to those observed with a threshold of 60 (slightly more contigs were obtained at 60 than 64 but with fewer YACs per contig). At the threshold of 64 the mean expected contig length is 1.54 Mb, 53% of the genome being covered with all the contigs and 22% with contigs larger than 3 Mb (Table 3). This is probably an

upper limit. The expected genome coverage c (in kilobases) of a given contig of n YACs can be estimated when considering 900 kb as the mean size of YACs in contigs and a 5-fold redundancy by the equation $c = 900 + (n - 1) \times 900/5$. Contigs larger than 3 Mb will then contain more than 13 YACs. From the observed data (Table 1), these contigs will cover 15% of the genome at a threshold of 60, and 22% at a threshold of 58. This number can be affected by YAC chimerism. Therefore, we estimate that 15% to 20% of the genome is covered with contigs larger than 3 Mb at a LOS threshold of 58.

Obviously, using only the L1 probe further mapping will progress very slowly (Figure 5), whereas employing a repeat probe that increases the information content of the L1-poor clones would be considerably more effective. A probe derived from the Alu consensus sequence, GM-007, was tested on 50 random YACs. This oligomer hybridizes only to the predicted variant and precise Alu subfamilies that constitute about 10% of the total Alu sequences (Matera et al., 1990). Under our conditions GM-007 produces an average of 15 fragments, much less than a total Alu consensus repeat, which is too frequent to be used for YAC fingerprinting. In Figure 1 hybridization patterns are shown on L1-poor and L1-negative clones. GM-007 gives an R banding pattern (G. Matera, personal communication) when hybridized on metaphase chromosomes. It is assumed that GM-007 in conjunction with the L1 probes will produce a set of contigs covering most of the human genome. When adding a new probe, the clone informativity increases and, therefore, the number of false positives for the same threshold will decrease. Simulations predict the same rate of false positives for the combination of the L1 and GM-007 probes at a threshold of 56, as for the L1 probe alone at a threshold of 64 (the Alu-like probe has been assumed to have a uniform distribution). Then according to the predictive model (Table 3 and Figure 5), when using both L1 and GM-007 probes on 46,000 clones, it is expected that 90% of the genome will be covered with contigs larger than 5 Mb.

Such large contigs will be easily mapped to metaphase chromosomes by FISH. It is still essential to develop an efficient ordering algorithm to facilitate the use of the map and to detect inconsistencies. The redundancy of fingerprints generated with three enzymes and two probes

Table 3. Expected Genome Coverage

| | Probe (LOS Threshold) | | | |
|-------------------------|-----------------------|---------|------------------|------------------|
| | L1 (64) | L1 (64) | L1 + GM-007 (56) | L1 + GM-007 (56) |
| Number of clones | 21,696 | 46,000 | 21,696 | 46,000 |
| Redundancy | 5 | 12.8 | 5 | 12.8 |
| Mean contig length (Mb) | 1.54 | 2.98 | 2.25 | 13.3 |
| Genome coverage with: | | | | |
| All contigs (%) | 53 | 87 | 85 | 95 |
| Contigs more than (%) | | | | |
| 2 Mb | 33 | 74 | 66 | 94 |
| 3 Mb | 22 | 64 | 52 | 93 |
| 5 Mb | 8.8 | 44 | 30 | 90 |
| 10 Mb | 0.6 | 13 | 5.4 | 78 |

should provide sufficient information to derive a reliable order. This map will constitute a backbone that will speed up the construction of higher resolution maps.

Experimental Procedures

Preparation of Total Yeast DNA

The library used was constructed as previously described (Albertsen et al., 1990) with a few modifications (P. O., unpublished data). To prepare DNA, 6 microtiter plates containing 160 μ l of AHC medium per well (Brownstein et al., 1989) were inoculated with a 96 needle replicator from the glycerol microtiter plate and grown at 30°C for 3 days to yield a concentration of 3.5×10^7 cells/ml. The cells were harvested by centrifugation, and cell pellets from the 6 microtiter plates were pooled and resuspended in 150 μ l of 1 M sorbitol, 10 mM EDTA, 0.1 M sodium citrate, 18 mM 2-mercaptoethanol, and 2 U of Zymolyase 20T (Kirin Brewery Co., Japan). The plates were incubated at 37°C for 20 min and centrifuged at 1300 rpm for 5 min. The spheroplast pellets were resuspended in 50 μ l of 25 mM Tris-HCl, 25 mM EDTA, 0.5% DLS, and 30 mM NaCl and incubated at 50°C with agitation for 2 hr. The lysates were transferred into a dialysis microtiter plate and dialyzed in a floating chamber against 10 mM Tris-HCl (pH 8), 0.1 mM EDTA (pH 8) at 55°C overnight. The resulting preparation corresponding to approximately 2 μ g of total yeast DNA can be stored for long periods at -20°C. The steps of pooling, addition, and resuspension were performed using the Hamilton robot (Hamilton, Switzerland).

YAC Digests and Southern Blot

YAC DNAs (approximately 300 ng) were digested with restriction endonucleases EcoRI, PstI, and PvuII (NBL) in 60 μ l reactions with NBL buffer (10 \times) plus 10 μ g/ml RNAase A (Boehringer Mannheim) at 37°C for 4 hr. The restriction digests were precipitated and resuspended in 5% Ficoll and 5% bromophenol blue 20 \times . Digested DNAs were loaded on a 0.8% agarose gel (Seakem LE, FMC Corp.), run in 0.5 \times TBE for 16 hr at 14°C, and transferred onto nylon membrane (Hybond N+, Amersham) for 30 min using an automated multiblotting device (Bertin, France). Eight gels of 15 slots (12 restriction digests and 3 size markers: a mixture of λ DNA digested with restriction enzymes XhoI, BssHII, or BstE2 and ϕ X174 digested with restriction enzymes HaeIII, MluI, BstNI, or NruI) corresponding to one 96-well microtiter plate of DNA, digested with one enzyme, were run per robot. Membranes were treated with 0.4 M NaOH and neutralized in 0.5 M Tris-HCl (pH 7.5), 1.5 M NaCl.

Hybridizations

The following were used as repetitive sequence probes: L1 probe, a 1.5 kb Kpn fragment from the 3' conserved L1 repetitive sequence (Shafit-Zagardo et al., 1982); GM-007, 5'-TGGATCAGAGGTGAGGAGGAGATCGAGACCATCCCGGCTAAACGGTGAA-3', located between positions 57 and 104 of a conserved Alu consensus sequence that detects only the predicted variant and precise Alu subfamilies (Matera et al., 1990). Two yeast single-copy probes were used: pAF001, a 1.5 kb EcoRI fragment from chromosome III (Thierry et al., 1990), and pAF020, a 1.8 kb EcoRI fragment from chromosome XI (Colleaux et al., 1992). DNA probes were labeled with horseradish peroxidase (ECL Direct system, Amersham), according to the manufacturer's instructions, directly or after a random elongation carried out by nucleotide incorporation using terminal deoxynucleotidyl transferase (Boehringer Mannheim) for the oligonucleotide probe. Membranes were hybridized for 16 hr at 42°C in ECL buffer supplemented with 0.5 M NaCl, 5% casein and washed at 55°C for 10 to 20 min in SSC at a concentration of 0.5 \times to 0.2 \times , depending on the probe. Membranes were detected with a chemiluminescent reaction carried out with luminol (Amersham) and visualized by photographic exposure for 30 to 60 min.

Inter-Alu PCR Fingerprint Analysis

The Alu primer used was the A33 primer 5'-CACTGCACTCCAAGCC-TGGGCGAC-3' (Chumakov et al., 1992). Inter-Alu PCR was carried out in a total volume of 30 μ l with 10 ng of YAC clone DNA, 10 ng/ μ l A33 primer, 250 μ M each dNTPs, 1.5 mM MgCl₂, 50 mM KCl, 10 mM Tris-HCl (pH 8.3), 0.1% Triton X-100, and 4 U of Taq polymerase

(NBL). Initial denaturation was for 5 min at 96°C, followed by addition of Taq polymerase at 92°C and 40 cycles of denaturation at 94°C, annealing at 60°C and extension at 72°C, each for 1 min, with a final extension at 76°C for 10 min. Inter-Alu PCR products were diluted in 80 μ l of 5% glycerol, 1 mM EDTA, and 0.01% xylene cyanol and run on a 2% agarose gel (FMC Corp.) containing ethidium bromide.

FISH

Metaphase chromosomes were prepared after methotrexate synchronization and bromodeoxyuridine incorporation as described previously (Arnold et al., 1992). Prior to in situ hybridization, slides were pre-treated with RNAase and pepsin, followed by a postfixation step in formaldehyde as described previously (Ried et al., 1992a).

FISH was carried out using inter-Alu PCR products of individual YAC clones or YAC contigs. Inter-Alu PCR products were labeled with biotin-11-dUTP by random priming (Feinberg and Vogelstein, 1983) in a 50 μ l reaction at 37°C for 1.5 hr. Excess nucleotides were removed with a Sephadex spin column. DNA probes were then digested with DNAase I to an average size of 150-400 bp.

For hybridization, DNA probes were precipitated and resuspended in 10 μ l of hybridization buffer (50% formamide, 2 \times SSC, 10% dextran sulfate), denatured at 75°C for 5 min, and allowed to preanneal at 37°C for 30 min. Probes were applied to the chromosome preparation, previously denatured at 80°C for 2 min in 70% formamide, 2 \times SSC, under a 18 \times 18 mm² coverslip. Hybridization, washings, and detection with avidin-fluorescein isothiocyanate (Vector Laboratories) were performed as previously described (Ried et al., 1992b). Signals were visualized on DAPI banded chromosomes using a Zeiss epifluorescence microscope and a cooled CCD camera (PM512, Photometrics).

Laboratory Notebook

A relational data base programmed in Sybase SQL, named CPFC, has been developed to serve as the laboratory notebook. It contains all the tables necessary to describe the fingerprinting process from colony picking to image analysis of Southern blots. This includes protocol descriptions, reagent batch numbers and dates, and experimental parameters.

Image Analysis

Films are scanned with a Truvel scanner driven by a Sun work station. The resulting digital images (260 dpi, 7.5 Mb/film) are written (via NFS) on two 1 gigabyte disks driven by two Sun Sparc 2 stations. Images are analyzed overnight: bands are detected and quantified as two-dimensional objects, then lanes are drawn and sizes assigned to standards. The molecular weights of detected fragments are computed from these standards and the results (molecular weight, area, maximum and integrated optical density for each band) are written into ASCII files. This automatic process, including scanning, does not exceed 4 min per image. Each image is edited to check the results, and modifications are performed if necessary; 70% of the images do not require any modification, and 20% need only minor correction. The treatment of the 10% remaining requires 10 to 20 min per image. Then, for each series, a program checks the quality of the results (consistency between the different enzymes, contaminations, and range of molecular weights and intensities). The operator is helped in his task by making queries to the CPFC data base (see above). Images and analysis files are moved to 600 Mb magneto-optical disks. The image editor and image analysis programs have been developed by Millipore-Biolmage. The checking, scheduling, and data manipulation programs were written internally. With this system, three persons were able to analyze more than 65,000 fingerprints (21,696 YACs, fingerprinted with three enzymes and one probe) within 4 months.

Data Analysis

Clone Comparisons

We performed pairwise comparisons, building the triangular matrix of overlap likelihoods. This likelihood ratio, i.e., the probability of observing clones with overlapping fingerprints versus nonoverlapping fingerprints, is computed from repetitive sequence fingerprints using Bayes theorem (Balding and Torney, 1991). These probabilities are obtained by summing up all possible ways of matching fragments (only fragments whose lengths are "close enough," according to an experimentally measured error, can be matched). The underlying model assumes

Gaussian error on fragment size measurements, with a standard deviation depending on fragment size, and models the distribution of both restriction sites and repetitive sequences with Poisson processes. Since the probability of observing clones with overlapping fingerprints is obtained by integration over all possible length of overlap, it is possible to determine the most probable overlap length. Further information can be found in Lacroix and Codani, 1992.

The amount of data to be treated ($C(C-1)/2$ comparisons, where C is the number of clones analyzed) leads to important computation times. For our current data (three independent fingerprints with L1, having an average of 5 fragments per clone), the computation for 20,000 clones takes 1 month of Sun Sparc Station 2 CPU time. For a more frequent probe (20 fragments per clone, for example), this duration could be increased to 2 years. Consequently, state of the art software techniques and hardware architectures have been used to exploit the intrinsic parallelism of the problem. As a result, computation for 20,000 clones has been performed in 1 day on a network of 20 heterogeneous Unix workstations. Computation can also be done incrementally at the rate of fingerprint production.

Contig Assembly

Once computed, the matrices are used by a simple thresholding algorithm to construct contigs for a given threshold. To determine a threshold that would give only a low number of false positives while not increasing the number of false negatives significantly, we used independently simulated data and 18,462 known nonoverlapping clone pairs from chromosome 21. Simulations on 10,000 clones indicated that a threshold of 55 would avoid false positives; when analyzing 21,696 clones, this threshold should give only a few false positives. From the LOS of the chromosome 21 nonoverlapping clone pairs, we intended to estimate the theoretical distribution of LOS for nonoverlapping pairs, and, therefore, to choose a threshold giving the tolerated rate of false positives, when assembling the 21,696 clones. But this requires a knowledge of the tail of the LOS distribution, i.e., an estimate with significantly more than 18,462 pairs. Therefore, the empirical curve was fitted in its asymptotic part by an adequate function (an Erlang function with parameter 2), and the integration of this function above any threshold gave the corresponding rate of false positives.

Predictive Model

The estimation of relative abundance of L1-rich and L1-poor regions was performed with a maximum likelihood test. The probability of observing the fingerprints of 953 clones of known length was maximized as a function of the relative abundance of the two regions (the frequencies of the repeat probe in both regions and the abundances being constrained to yield the observed mean frequency).

Each clone was assumed either to have a fixed minimum detectable overlap or to be unmatchable for lack of information. The density of minimal detectable overlap and the proportion of unmatchable clones was computed from the simulated data, as a function of the clone size. The chimera likelihood was presumed to be independent of clone size and only the larger insert of a chimera was admitted for matching. Derivations of the equations giving the contig characteristics are similar to Lander and Waterman, 1988.

Acknowledgments

We would like to thank Claude Scarpelli for assistance in computer system management and Greg Matera for providing the GM-007 oligonucleotide. This work was supported by the Ministère de la Recherche et de la Technologie and by the Association Française contre les Myopathies. C. B.-C. was supported by a fellowship from the Institut National de Santé et de la Recherche Médicale and B. L. was supported by a fellowship from the Ecole Polytechnique.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC Section 1734 solely to indicate this fact.

Received August 26, 1992; revised September 2, 1992.

References

- Albertsen, H. M., Abderrahim, H., Cann, H. M., Dausset, J., Le Paslier, D., and Cohen, D. (1990). Construction and characterization of a yeast artificial chromosome library containing seven haploid genome equivalents. *Proc. Natl. Acad. Sci. USA* 87, 4256-4260.
- Anand, R., Villasente, A., and Tyler-Smith, C. (1989). Construction of yeast artificial libraries with large inserts using fractionation by pulsed-field gel electrophoresis. *Nucl. Acids Res.* 17, 4325-4333.
- Arnold, N., Bhatt, N., Ried, T., Ward, D. C., and Wienberg, J. (1992). Fluorescence in situ hybridization on banded chromosomes. In *Techniques and Methods in Molecular Biology: Non-Radioactive Labeling and Detection of Biomolecules*, C. Kessler, ed. (Berlin/Heidelberg/New York: Springer Verlag), in press.
- Balding, D. J., and Torney, D. C. (1991). Statistical analysis of DNA fingerprint data for ordered clone physical mapping of human chromosomes. *Bull. Math. Biol.* 53, 853-879.
- Bates, G. P., Valdes, J., Hummerich, H., Baxendale, S., Le Paslier, D. L., Monaco, A. P., Tagle, D., MacDonald, M. E., Altherr, M., Ross, M., Brownstein, B. H., Bently, D., Wasmuth, J. J., Gusella, J. F., Cohen, D., Collins, F., and Lehrach, H. (1992). Characterization of a yeast artificial chromosome contig spanning the Huntington's disease gene candidate region. *Nature Genet.* 1, 180-187.
- Bellanné-Chantelot, C., Barillot, E., Lacroix, B., Le Paslier, D., and Cohen, D. (1991). A test case for physical mapping of human genome by repetitive sequence fingerprints: construction of a physical map of a 420 kb YAC subcloned into cosmids. *Nucl. Acids Res.* 19, 505-510.
- Botstein, D., White, R. L., Skolnick, M., and Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* 32, 314-331.
- Brownstein, B. H., Silverman, G. A., Little, R. D., Burke, D. T., Korsmeyer, S. J., Schlessinger, D., and Olson, M. V. (1989). Isolation of single-copy human genes from a library of yeast artificial chromosome clones. *Science* 244, 1348-1351.
- Burke, D. T., Carle, G. F., and Olson, M. V. (1987). Cloning of large exogenous DNA into yeast by means of artificial chromosome vectors. *Science* 236, 806-812.
- Cangiano, G., Ameer, H., Waterston, R., and La Volpe, A. (1990). Use of repetitive DNA probes as physical mapping strategy in *Caenorhabditis elegans*. *Nucl. Acids Res.* 18, 5077-5081.
- Chumakov, I. M., Le Gall, I., Billault, A., Ougen, P., Soularue, P., Guillou, S., Rigault, P., Bui, H., De Tand, M. F., Barillot, E., Abderrahim, H., Cherif, D., Berger, R., Le Paslier, D., and Cohen, D. (1992). Isolation of chromosome 21-specific yeast artificial chromosomes from a total human genome library. *Nature Genet.* 1, 222-225.
- Colleaux, L., Richard, G. F., Thierry, A., and Dujon, B. (1992). Sequence of a segment of yeast chromosome XI identifies a new mitochondrial carrier, a new member of the G protein family and a protein with the PAAKK motif of the H1 histones. *Yeast* 8, 325-336.
- Coulson, A., Sulston, J., Brenner, S., and Kam, J. (1986). Towards a physical map of the nematode *C. elegans*. *Proc. Natl. Acad. Sci. USA* 83, 7821-7825.
- Cox, D. R., Burmeister, M., Price, E. R., Kim, S., and Myers, R. M. (1990). Radiation hybrid mapping: a somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes. *Science* 250, 245-250.
- Craig, G., Nizetic, D., Hoheisel, J. D., Zehetner, G., and Lehrach, H. (1990). Ordering of cosmid clones covering the Herpes simplex virus type 1 (HSV 1) genome: a test case for fingerprinting by hybridization. *Nucl. Acids Res.* 18, 2653-2660.
- Evans, G. A., and Lewis, K. A. (1989). Physical mapping of complex genomes by cosmid multiplex analysis. *Proc. Natl. Acad. Sci. USA* 86, 5030-5034.
- Feinberg, A. P., and Vogelstein, B. (1983). A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. *Anal. Biochem.* 132, 6-13.
- Green, E. D., Riethman, H. C., Dutchik, J. E., and Olson, M. V. (1991). Detection and characterization of chimeric yeast artificial-chromosome clones. *Genomics* 11, 658-669.
- Jeffreys, A. (1987). Highly variable minisatellites and DNA fingerprint. *Biochem. Soc. Trans.* 15, 309-317.
- Kohara, Y., Akiyama, K., and Isono, K. (1987). The physical map of the whole *E. coli* chromosome: application of a new strategy for rapid

- analysis and sorting of a large genomic library. *Cell* 50, 495-508.
- Korenberg, J. R., and Rykowski, M. C. (1988). Human genome organization: Alu, lines, and the molecular structure of metaphase chromosome bands. *Cell* 53, 391-400.
- Korenberg, J. R., Yang-Feng, T., Schreck, R., and Chen, X. N. (1992). Using fluorescence in situ hybridization (FISH) in genome mapping. *Trends Biotech.* 10, 27-32.
- Lacroix, B., and Codani, J. J. (1992). Proceedings of the second international conference on bioinformatics, supercomputing and complex genome analysis. World Sci., in press.
- Lander, E. S., and Waterman, M. S. (1988). Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* 4, 231-239.
- Larin, Z., Monaco, A. P., and Lehrach, H. (1991). Yeast artificial chromosome libraries containing large inserts from mouse and human DNA. *Proc. Natl. Acad. Sci. USA* 88, 4123-4127.
- Matera, G. A., Hellmann, U., Hintz, M. F., and Schmid, C. W. (1990). Recently transposed Alu repeats from multiple source genes. *Nucl. Acids Res.* 18, 6019-6023.
- Montanaro, V., Casamassimi, A., D'Urso, M., Yoon, J. Y., Freije, W., Schlessinger, D., Muenke, M., Nussbaum, R. L., Saccone, S., Mauri, S., Santoro, A. M., Motta, S., and Valle, G. D. (1991). In situ hybridization to cytogenetic bands of yeast artificial chromosomes covering 50% of human Xq24-Xq28 DNA. *Am. J. Hum. Genet.* 48, 183-194.
- Olson, M., Dutchik, J. E., Graham, M. Y., Brodeur, G. M., Helms, C., Frank, M., Mac Collin, M., Scheinman, R., and Frank, T. (1986). Random clone strategy for genomic restriction mapping in yeast. *Proc. Natl. Acad. Sci. USA* 83, 7826-7830.
- Ried, T., Lengauer, C., Cremer, T., Wiegant, J., Raap, A. K., van der Ploeg, M., Groitl, P., and Lipp, M. (1992a). Specific metaphase and interphase detection of the breakpoint region in 8q24 of Burkitt lymphoma cells by triple color fluorescence in situ hybridization. *Genes Chromosom. Cancer* 4, 69-74.
- Ried, T., Baldini, A., Rand, T. C., and Ward, D. C. (1992b). Simultaneous visualization of seven different DNA probes by in situ hybridization using combinatorial fluorescence and digital imaging microscopy. *Proc. Natl. Acad. Sci. USA* 89, 1388-1392.
- Shafit-Zagardo, B., Maio, J. J., and Brown, F. L. (1982). L1 families of long, interspersed repetitive sequences in human and other primate genomes. *Nucl. Acids Res.* 10, 3175-3193.
- Stallings, R. L., Torney, D. C., Hildebrand, C. E., Longmire, J. L., Deaven, L. L., Jett, J. H., Dogget, N. A., and Moysis, R. K. (1990). Physical mapping of human chromosomes by repetitive sequence fingerprinting. *Proc. Natl. Acad. Sci. USA* 87, 6218-6222.
- Thierry, A., Fairhead, C., and Dujon, B. (1990). The complete sequence of the 8.2 kb segment left of MAT on chromosome III reveals five ORFs, including a gene for a yeast ribokinase. *Yeast* 6, 521-534.